METHOD AND SYSTEM FOR CLASSIFYING INFORMATION AVAILABLE ON A COMPUTER NETWORK

Background of the Invention

The present invention relates to a system and method for monitoring and analyzing computer network transaction data to create behavior profiles of network users. More particularly, the present invention relates to a method and system of manually or automatically classifying information available on a computer network. Specifically, the present invention helps classify Internet Web sites to facilitate the construction of more accurate behavior profiles of Internet users for marketing purposes.

In the current Internet world, it has become desirable for service providers and merchants to obtain specific information about Internet users for the purpose of improving the marketing of products and services, measuring the effectiveness of marketing, and tailoring the products and services to meet the requirements of specific customer types.

Behavior profiles are created using network usage data collected through various methods. Once the data is collected, it is analyzed to determine the behavior of a particular user. In order to create an accurate behavior profile, it is useful to generalize Internet usage by identifying the types of Web sites a particular type of user accesses and the way that type of user accesses a particular type of Web site.

For example, it would be valuable to a merchant to know that users from a geographical area regularly purchase books from Amazon.com™; however, there is a need for more generalized data. It is desirable to have a system that can create generalized behavior profiles. It is valuable information to know that users in a particular geographical

area regularly conduct electronic commerce by accessing online catalog and shopping sites by following links on a Web portal site.

To provide a system for creating generalized behavior profiles, it is desirable to have a method and system for classifying Web sites using a classification of sufficient granularity to allow meaningful analysis of network transaction data.

Manual classification by users can lead to inconsistent results due to differing understandings of categories within a classification system, differing opinions of the purpose and use of a site, etc. It is desirable to have a method and system that provides a more consistent categorization of information. Additionally, it is desirable to provide a system and method such that inexperienced classifiers can perform the bulk of classification without sacrificing accuracy.

Also, there is a need for an automatic classification system that can quickly and accurately categorize information repositories accessible on a computer network. An automatic classification system can operate more quickly and at less expense than a manual classification system; however, the automatic classification system may not be as accurate as a manual classification system.

Finally, there is a need for a hybrid classification system that uses both manual and automatic classification components to provide increased performance and accuracy.

Summary of the Invention

In accordance with the invention, there is provided a method for classifying information available on a computer network. The method includes receiving a list of

network resource locators. For each network resource locator on the received list, the method includes sending the network resource locator to a Web-coding workstation. Once the locator has been sent to the workstation, the process waits for a vote to be received from the Web coder. Each vote represents a proposed classification for a network resource locator. The result is stored in a database. Finally, the process assigns a classification according to a voting system. In a more specific embodiment, the list of network resource locators includes one or more Web sites. In a further aspect of the invention, the database is a flat file, a binary tree, an object-oriented database, or a relational database.

Additional, more specific embodiments include various voting methods. First, a single-level voting system is presented wherein a classification is assigned upon receipt of a single vote. Next, a single-level voting system is presented requiring more than one vote. Finally, a multiple-level voting system is presented wherein the first level requires three out of four votes, the second level requires two out of three votes, and the third level requires a single vote.

Further in accordance with the present invention, there is provided a system for classifying information available on a computer network. The system includes a resource generator component that creates a list of network resource locators. Additionally, the system includes a datastore component that stores classification information for a plurality of network resource locators. More specific embodiments use a flat file, an object-oriented database, a binary tree, or a relational database as the datastore. The system also includes a graphical user interface (GUI) component and a classification processor component. The

classification processor receives a list of network resource locators and determines a classification for each information repository.

More specific embodiments include a system that sorts the list of network resource locators by the number of unique visitors to that particular locator. Additional embodiments use the various voting systems set forth in the method above.

Brief Description of the Drawings

Having thus briefly described the invention, the same will become better understood from the following detailed discussion, taken in conjunction with the drawings where:

Figure 1 is a general system schematic diagram showing a coding workstation connected to a computer network and the Internet so that the coding workstation has access to the Web server implementing the present invention and a plurality of information repositories;

Figure 2 is a schematic diagram illustrating the interactions between the various components according to one embodiment of the present invention;

Figure 3A is a flow chart illustrating how one embodiment of the present invention obtains a list of network resources, sorts and prioritizes the list, and then classifies each of the identified network resources;

Figure 3B is a flow chart illustrating how the classification-processing component interacts with a database and a Web coder to classify a network resource according to one embodiment of the present invention; and

Figure 4 depicts a sample screen display that a Web coder would see on a coding workstation according to one embodiment of the present invention.

Detailed Discussion of the Invention

The present invention includes methods and systems for manually and automatically classifying information on a computer network. In order to simplify explanations of the invention, the following detailed discussion is limited to classifying Web sites on the Internet; however, one of ordinary skill in the art will understand that the invention is not so limited; the method and system herein described may be used to classify any data made available on any network.

The first embodiment of the present invention is a single-vote classification system implemented as a Web-based application using conventional software development techniques familiar to one of ordinary skill in the art. The single-vote system runs on a Web server 101 embodied as a Sun Microsystems™ Enterprise 6500™ server running Apache™ Web server software.

Figure 1 is a diagram of a hardware implementation of the present invention. The Web server 101 includes a conventional network interface 102 that connects the Web server 101 to a computer network 103. In this embodiment of the present invention, computer network 103 is also connected to the Internet 104. A coding workstation 105 also connects to the computer network 103 using a conventional network interface 106.

Figure 2 is a block diagram of the software implementation of this embodiment of the present invention. In this embodiment, all of the server software runs on the Web server 101.

According to the present invention, network information repositories are categorized according to a classification system. The first embodiment uses a manual classification process to go through and classify various identified Web sites. The list of Web sites to be classified is generated by the resource generator 201.

In one embodiment of the present invention, the resource generator 201 is a Standard Query Language (SQL) statement that directly calls the database to retrieve a list of network resource locators in rank order. This list is used to populate another database table. When the number of unclassified hosts gets low, the resource generator 201 makes another SQL call to retrieve additional network resource locators and populates the database table accordingly.

In another embodiment of the present invention, the resource generator 201 simply inputs a text file containing a list of Web site addresses stored as a list of Uniform Resource Locators (URL). For example, the list may contain the following locators: "http://www.amazon.com" and "http://www.bn.com". In this embodiment, the resource generator 201 processes the list of URLs stored in the text file in order, passing each one to the next component of this embodiment--a classification processor 202.

In an alternate embodiment of the present invention, the classification system is used with a network transaction data-collection system. In this embodiment, the network transaction data-collection system maintains data and statistics on Internet or other computer network usage. The collected network transaction data is used to generate a list of all network resources accessed by users. In most applications, this list quickly becomes unwieldy, so it is desirable to prioritize the list, pruning it down to something manageable.

In one embodiment of the present invention, the network transaction data-collection system maintains a list of servers on the Internet that have been accessed by network users. Along with each server, the system maintains the number of unique users that have accessed that server. The list of servers is then sorted based on the number of unique users. This determines the order that servers are classified.

One of ordinary skill in the art will appreciate that there are many other ways to create a list of network resources that need to be categorized. Any such method or system could be used within the present invention.

Once one or more resources to be classified has been determined by the resource generator 201, the classifying processor 202 begins. The classifying processor 202 presents a resource to a user through the GUI 203. The user then looks at the Web site and identifies the classification of that Web site as discussed below.

Web sites are classified according to a system loosely based on the North American Industry Classification System (NAICS) that replaced the Standard Industry Classification (SIC) codes in 1997. NAICS was created to provide a way to compare statistics regarding businesses in the same industry. While providing adequate granularity for describing and classifying all types of businesses within the United States, NAICS fails to provide sufficient granularity for the domain where the present invention is most useful--electronic commerce.

The electronic commerce classification system (ECCS) used in the present invention includes hierarchically arranged categories including classifications such as Web portals and online catalogs.

In one embodiment of the present invention, a Web coder accesses the Web server 101 using a Web browser such as Microsoft Internet Explorer™ or Netscape Communicator™ running on a coding workstation 105 to view and categorize Web pages. In this embodiment, a Web coder logs on to the Web server 101 by typing a username and password. The system authenticates the user and then displays a list of various statistics about that user, including the total number of Web sites classified, the total classified in the present week, and the total classified during the present day.

The Web site is implemented using conventional software development techniques known to one of ordinary skill in the art. The system runs on an Apache™ Web server as a series of Common Gateway Interface (CGI) scripts written in Perl™. These scripts present a frame-based GUI to the present Web classification system. At the top of the screen is a main frame 401 situated horizontally across the Web browser. This frame displays a logo identifying the Web classification system, a series of statistics about the current authenticated Web coder, and a button labeled "Retrieve Resource". When the user presses the "Retrieve Resource" button, the next available resource is retrieved from the resource generator 201.

While a resource is being displayed by the classification system, the main frame 401 remains displayed, and the resource is shown in the browser frame 402, a frame situated below the main frame 401 and to the right of a category frame 403. The Web coder can browse the displayed resource in the browser frame 402, following links and examining the content of the resource site. The user then selects a resource from the hierarchical taxonomy displayed in the category frame 403.

Each category display in category frame 403 is either a parent category or a terminal category. For example, "electronic commerce shopping site" would be a parent category including child categories such as "music", "books", and "computers." The hierarchical structure is displayed in the conventional manner, allowing a user to compress and expand various nodes within the structure.

Sub

Once a Web coder has determined the appropriate classification, the classifier processor 202 updates a database containing a list of all classified network resources. The system also updates the statistics for that user and displays the changed values in the main frame 401. The user can then select the "Retrieve Resource" button again to obtain the next Web site to be classified. This process repeats until the user chooses to stop or the resource generator 201 runs out of servers to classify.

This embodiment of the present invention can encounter problems with Web sites designed to eliminate extraneous frames. According to another embodiment of the present invention, the category frame 403 is implemented as a pull-down menu in which a coder can select the category that best matches the Web page being classified. This embodiment displays a URL. When a coder clicks on the URL, the Web site to be classified opens up in a another window. This prevents that Web site from interfering with the present classification system.

This embodiment of the present invention can support multiple Web coders working to classify a series of Web sites. The resource generator 201 can work in one of several different ways: (1) it can generate only a single resource that needs to be classified at a time; or (2) it can generate a predetermined number for each Web coder and then the classifying

processor 202 can process the block. When the block has been completed, the resource generator 201 transmits a new block.

Sometimes, a Web coder's classification may not be accurate. This may be due to several reasons. For example, the Web coder may be inexperienced and somewhat unfamiliar with the particular portion of the classification hierarchy that is relevant; or, the Web coder may misinterpret the purpose behind the site due to the limited time that the Web coder took to view that particular site.

Sub A6

Another embodiment of the present invention improves the accuracy of the classification system by implementing a voting process. Instead of using a single Web coder to classify a given Web site, the system gives queries to at least three different Web coders before accepting an identified classification. Realizing that there will be a difference of opinion as to the classification of some Web sites, the system does not require a unanimous consensus, instead using a multi-level voting system.

At the lowest level are the Level 1 coders. These are typically newer, less experienced classification specialists. At this first level, the system requires at least three out of four coders to agree before a final classification is accepted. If fewer than three out of four agree on a classification, the Web site is passed to the Level 2 coders.

Level 2 coders have more experience and knowledge about the classification system and are able to determine a classification with greater accuracy. At Level 2, two out of three coders must agree on a classification for it to be accepted by the system.

Finally, the top-level coders are called Expert Coders. These individuals usually have the greatest understanding of the classification system. Whatever classification a Expert Coder gives is accepted by the system.

Thus, most classification will be performed by Level 1 coders. Any confusion or disagreements over the appropriate classification will be passed on to a smaller number of Level 2 coders. Finally, a Expert Coder has the ultimate authority to determine a final classification. As part of the status information displayed in the main frame 401, the system displays the votes placed by lower-level coders.

The next component of the system is an automatic classification agent. Such an agent determines the appropriate classification of a Web site without any input from a user. Since the system is susceptible to error, an automatic classification agent counts as a Level 1 coder vote in the multi-level voting system discussed above. By using an automatic classification agent, fewer Level 1 coders are needed to maintain the level of accuracy.

Embodiments of the present invention have now been generally described in a non-limiting manner. It will be appreciated that these examples are merely illustrative of the present invention, which is defined by the following claims. Many variations and modifications will be apparent to those of ordinary skill in the art.